Random thresholds for linear model selection

Marc Lavielle, Carenne Ludeña

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Random thresholds for linear model selection

Marc Lavielle — Carenne Ludeña

N° 5572

Mai 2005

Thème COG

ISSN 0249-6399 ISRN INRIA/RR--5572--FR+ENG

Random thresholds for linear model selection

Marc Lavielle * , Carenne Ludeña †

Thème COG — Systèmes cognitifs

Projet Select

Rapport de recherche n° 5572 — Mai 2005 — 23 pages

Abstract: A method is introduced to estimate the number of significant coefficients in non ordered model selection problems. The method is based on a convenient random centering of the partial sums of the ordered observations. Based on $L-$statistics methods we show consistency of the proposed estimator. An extension to unknown parametric distributions is considered. The method is then applied to a regression model and interpreted as a random threshold procedure. Simulated examples are included to show the accuracy of the estimator.

Key-words: adaptive estimation, linear model selection, hard thresholding, random thresholding, L statistics.

This work was supported by ECOS Nord V00M03

* Université René Descartes and Université Paris-Sud, France. e-mail: Marc.Lavielle@math.u-psud.fr
† IVIC, Venezuela E-mail: cludena@ivic.ve

Unité de recherche INRIA Futurs
Parc Club Orsay Université, ZAC des Vignes,
4, rue Jacques Monod, 91893 ORSAY Cedex (France)
Téléphone : +33 1 72 92 59 00 — Télécopie : +33 1 72 92 59 ??

Sélection de modèles linéaires par seuillage aléatoire

Résumé : Une nouvelle méthode est proposée pour estimer le nombre de coefficients significatifs dans un problème de sélection de modèles. Cette méthode utilise un centrage aléatoire bien choisi des sommes cumulées partielles des observations ordonnées. En utilisant des propriétés des L-statistiques, nous montrons la consistance de l'estimateur proposé. Une extension à des distributions paramétriques inconnues est considérée. La procédure est ensuite appliquée à un modèle de régression et est interprétée comme une procédure de seuillage aléatoire. Des exemples numériques illustrent l'intérêt pratique de la méthode.

Mots-clés : estimation adaptative, sélection de modèle linéaire, seuillage dur, seuillage aléatoire, L-statistique.

1 Introduction

Consider the following model

$$y_i = \mu_i + \varepsilon_i,\ i = 1, \ldots, n$$

where (μ_i) is an unknown sequence of constants some of which are zero and (ε_i) are independent random variables with common cumulative distribution F_ε. The problem we study in this article is choosing the significant, non zero, coefficients based on the observations (y_i). Obviously, significant coefficients will be those which are greater than a certain threshold, i.e. choose index i if

$$|y_i| > \tau. \tag{1}$$

The choice of τ in (1) will depend on the distribution of the sequence (ε_i). So in practice τ must be calibrated in terms of the data. A usual technique is to consider a sequence of thresholds (τ_j) for values ranging from very small (many significant coefficients) to very big (few significant coefficients) and study the point where a substantial decrease in the number of significant coefficients occurs. Of course choosing the "right" τ is equivalent to choosing the "right" number k of significant coefficients, with the advantage that this can be done independently of the choice of (τ_j) by looking at the relative size of the observations. Indeed, a jump in the relative size of the observations should indicate the existence of significant (not noise) coefficients.

This has been considered by a number of authors (see, for example, [6, 7, 10]) and many adaptive procedures aimed at studying the correct "jump point" have been developed.

A natural approach seems to consider the partial sums of the absolute (or squared) observations ordered decreasingly, and study the fluctuations of these partial sums around some kind of centering factor.

In this article we tackle this problem based on the use of order statistics by considering a convenient random centering: the conditional expectation, with respect to the total sum, of these partial sums. Even when this conditional expectation cannot be computed in a closed form, an exponential change of variable makes this centering possible. We then construct an $L-$ statistic and study its weak convergence. Empirical probability tables or simulated ones based on the limiting process can then be constructed to accept or reject the null hypothesis of all coefficients being equal to zero. If we reject the null hypothesis, further inspection of the test statistic yields the subset of significant coefficients. Indeed, we construct a test statistic based on the minimization of a certain functional of the conveniently centered partial sums and show consistency of the estimated number of significant coefficients under mild assumptions over the gap between significant and non significant coefficients.

The above method requires previous knowledge of F_ε. In a parametric setting, $F_\varepsilon = F_\varepsilon(\,\cdot\,;\theta^\star)$, we show that the proposed method can also address the case $\theta^\star$ unknown, assuming the existence of a consistent estimator of $\theta^\star$. When $\theta^\star$ is a scale parameter, an appropriate modification of the estimating procedure yields a scale free statistic, which is also shown to be consistent.

We then apply our method to the problem of estimating the number of significant coefficients for the regression problem

$$y_i = f(x_i) + \eta_i,\ i = 1, \ldots, n$$

where f is an unknown function in some function space S and η_i are independent random variables with variance σ^2. A usual estimation procedure is to consider $f \in L^2(\mu)$ and a finite orthonormal system $\{\phi_\lambda\}_\Lambda$, with $|\Lambda| = M_n$. Denoting by $\langle y, \phi_j \rangle_n = 1/n \sum_{i=1}^n y_i \phi_\lambda(x_i)$ the empirical coefficients, Donoho and Johnstone [5] in their seminal article proposed choosing only those coefficients whose absolute value exceeded a certain threshold $u = \sqrt{\frac{\tau \sigma^2 \log(n)}{n}}$. This procedure has since been refereed to as hard thresholding.

In a very interesting reinterpretation, Barron, Birge and Massart [2] study the problem of hard thresholding in the context of non ordered model selection based on the addition of a penalization term. Their arguments are combinatorial based on the complexity of the underlying linear spaces: the size of the set of all possible models of size k out of K is bounded by $(eK/k)^k$ and a logarithmic factor depending on K must be introduced in order to bound the probabilities. In terms of (1) our observations would now be the empirical coefficients $\langle y, \phi_j \rangle_n$. Of course, except for the case $\eta_i \sim N(0, \sigma^2)$, the empirical coefficients will not be necessarily independent, although uncorrelated, so that the problem does not comply to our assumptions. However, in practice the method works well. As discussed in section 4.3, our method can be interpreted as a random threshold procedure .

The article is organized as follows: in section 2 we introduce the problem and basic notation as well as the proposed test procedure. In section 3 we state and prove theoretical results that justify our procedure, namely consistency of the selected subset of significant coefficients. In section 4.3 we consider certain extensions which include the parametric distribution case, an application to the problem of non ordered linear model selection for the regression setting and interpret out testing scheme in terms of a random penalization procedure. In section 5 we present simulated examples.

2 Describing the procedure

2.1 A first hypothesis testing procedure

Assume we observe $y_i = \mu_i + \varepsilon_i$. Variables ε_i are assumed to be independent and identically distributed with common cumulative distribution F_ε. We begin by assuming that the cumulative distribution function $F_{|\varepsilon|}$ of the $|\varepsilon_i|$'s is known. In section 4 we will deal with the unknown $F_{|\varepsilon|}$ case.

Given the collection $(y_i; 1 \leq i \leq n)$, we are interested in this section in testing if all the μ_i's are null or not. Thus, we introduce the following hypothesis:

Null hypothesis:

$\boldsymbol{H_0}$: $\mu_i \equiv 0$ for $i = 1, \ldots, n$.

Alternative hypothesis:

$\boldsymbol{H_1}$: there exists a non empty subset I of $\{1, 2, \ldots, n\}$ such that $\mu_i \neq 0$ for $i \in I$.

Then, the test procedure is defined as follows:

i) Order the observations $|y_{(1)}| \geq |y_{(2)}| \geq \ldots \geq |y_{(n)}|$

ii) For $i = 1, \ldots, n$, let $X_{(i)} = -\log\left(1 - F_{|\varepsilon|}(|y_{(i)}|)\right)$,

iii) Let $T_j = \sum_{i=1}^{j} X_{(i)}$ and $Q_j = \mathbb{E}_{H_0}\left(T_j | T_n\right)$.

iv) Define the test statistic $D_n = \max_j |T_j - Q_j|/\sqrt{n}$. We will reject the null hypothesis if $D_n > d_\alpha$, where d_α is defined in Section 3.

Remark 1: Under the null hypothesis, the sequence $(X_{(i)})$ is a decreasing sequence of exponential random variables with parameter 1. Then, the conditional expectation $\mathbb{E}_{H_0}\left(T_j | T_n\right)$ can easily be computed using the following proposition (the proof is given in the appendix):

Proposition 2.1 *Assume $X_{(1)}$, $X_{(2)}$, ... , $X_{(n)}$ is an ordered sequence of Exp(1) random variables, with $X_{(1)} \geq X_{(2)} \geq \ldots X_{(n)}$. For any $1 \leq j \leq n$, let $T_j = \sum_{i=1}^{j} X_{(i)}$. Then, for any $j \leq K \leq n$,*

$$\mathbb{E}\left(X_{(i)}\right) = \sum_{\ell=1}^{n} \frac{1}{\ell} \tag{2}$$

$$\mathbb{E}\left(T_j\right) = j + j \sum_{i=j+1}^{n} \frac{1}{i} \tag{3}$$

$$\mathbb{E}\left(T_j | T_K\right) = \frac{\mathbb{E}\left(T_j\right)}{\mathbb{E}\left(T_K\right)} T_K. \tag{4}$$

Remark 2: The distribution of the test statistic D_n cannot be computed in a closed form. Nevertheless, the following standard result will allow us to construct probability tables (the proof is given in the Appendix):

Theorem 2.1 *Assume $X_{(1)}$, $X_{(2)}$, ... , $X_{(n)}$ is an ordered sequence of Exp(1) random variables, with $X_{(1)} \geq X_{(2)} \geq \ldots X_{(n)}$. For any $1 \leq j \leq n$, let $T_j = \sum_{i=1}^{j} X_{(i)}$. Introduce for $t \in [0,1]$ the random process $d_n(t) = T_{[nt]} - \mathbb{E}\left(T_{[nt]} | T_n\right)$. Then, $\frac{1}{\sqrt{n}} d_n(t)$, as a stochastic process indexed on $t \in [0,1]$, converges in distribution to a zero mean Gaussian process Δ with covariance function defined by*

$$\begin{aligned}\mathbb{E}\left(\Delta(t)\Delta(s)\right) = & \int_0^1 \int_0^1 [(1-u) \wedge (1-v) - (1-u)(1-v)][\mathbb{1}_{[0,t]}(u) - t + t\log(t)] \\ & \times [\mathbb{1}_{[0,s]}(v) - s + s\log(s)] dG^{-1}(u) dG^{-1}(v),\end{aligned}$$

where $G(x)$ is the distribution function of an exponential r.v.

Using Theorem 2.1, we can conclude that statistic D_n defined in the test procedure converges weakly to $\Delta_\infty = \sup_t \Delta(t)$. Then, d_α is defined as the α quantile of Δ_∞.

Remark 3: Instead of assuming that the distribution of the $|\varepsilon_i|$ is known, we can assume that there exists an increasing continuous function $h : \mathbb{R}^+ \to \mathbb{R}^+$ such that the cumulative distribution function F_h of $h(|\varepsilon|)$ is known. Then, $X_{(i)}$ is defined as $-\log\left(1 - F_h(h(|y_{(i)}|))\right)$. Without any loss of generality, we will consider the case $h = id$ in the following.

Remark 4: A uniform change of variable can also be used, by setting $X_{(i)} = F_{|\varepsilon|}(|y_{(i)}|)$. Indeed, the conditional expectation of T_j can also be computed here:

$$\mathbb{E}\left(T_j|T_K\right) = \frac{j(K-j)}{K+1} + \frac{j(j+1)}{K(K+1)}\,T_K$$

2.2 Choosing the right coefficients

If we reject the null hypothesis, the next step is to select the significant coefficients.

Let s be the one to one mapping from $\{1,2,\ldots,n\}$ to $\{1,2,\ldots,n\}$ defined by $Y_{s(i)} = Y_{(i)}$ (recall that $(Y_{(i)})$ is a decreasing sequence).

Then, define the set of alternative hypotheses:

Alternative hypothesis:

$\boldsymbol{H_1(k)}$: there exists a subset $I_k \subset \{1,2,\ldots,n\}$ such that,
- for any $i \in I_k$, $s(i) \leq k$ and $\mathbb{E}_{H_1(k)}\left(Y_i\right) \neq 0$,
- for any $i \notin I_k$, $s(i) \geq k+1$ and $\mathbb{E}_{H_1(k)}\left(Y_i\right) = 0$,

Under $\boldsymbol{H_1(k)}$, there are k significant coefficients and $|y_{(k+1)}|,\ldots,|y_{(n)}|$ have distribution $F_{|\varepsilon|}$. Then, we define the following test procedure:

i) For $i = 1,\ldots,n$, let $X_{(i)} = -\log\left(1 - F_{|\varepsilon|}(|y_{(i)}|)\right)$,

ii) Let K_n be some positive integer. For $1 \leq k \leq n - K_n$ and $1 \leq j \leq K_n$, compute

$$T_{k,j} = \sum_{i=k+1}^{k+j} X_{(i)}, \tag{5}$$

$$Q_{k,j} = \mathbb{E}_{H_1(k)}\left(T_{k,j}|T_{k,K_n}\right), \tag{6}$$

$$\eta_k = \max_{1\leq j\leq K_n} \frac{|T_{k,j} - Q_{k,j}|}{\sqrt{n}}. \tag{7}$$

iii) Let

$$\hat{k} = \mathrm{Arg}\min_{1\leq k\leq n-K_n} \eta_k$$

Remark 1: The ℓ_1 or the ℓ_2 norms can be used instead of the ℓ_∞ norm to define η by setting

$$\eta_k = n^{-\frac{3}{2}} \sum_{j=1}^{K_n} |T_{k,j} - Q_{k,j}|$$

or

$$\eta_k = n^{-2} \sum_{j=1}^{K_n} (T_{k,j} - Q_{k,j})^2$$

Remark 2: $Q_{k,j}$ can easily be computed using the results of the previous section. Indeed, Let

$$B_{k,j,n} = \frac{\mathbb{E}_k\left(T_{k,j}\right)}{\mathbb{E}_k\left(T_{k,K_n}\right)} = \frac{j\left(1+\sum_{i=j+1}^{n-k} 1/i\right)}{K_n\left(1+\sum_{i=K_n+1}^{n-k} 1/i\right)} \tag{8}$$

Then, Proposition 2.1 yields $Q_{k,j} = B_{k,j,n}\, T_{k,K_n}$.

In order to state our main consistency result, we consider the following asymptotic framework:

AF1 There exists $t^\star \in (0,1)$ and a subset $I_{k_n^\star}$ of $\{1,2,\ldots,n\}$ with $k_n^\star = [t^\star n]$, such that $\mu_i \neq 0$ if $i \in I_{k_n^\star}$. For all other index, $\mu_i = 0$.

AF2 For any $i \in I_{k_n^\star}$, $|\mu_i| \geq \alpha_n$, where $\alpha_n \to \infty$ according to the distribution of the (ε_i). Let $\Phi_{(1)}$ be the distribution of $\max_{1\leq i\leq n}|\varepsilon_i|$ and (a_n, b_n) such that $\Phi_{(1)}(a_n + b_n x) \to W(x)$ for some fixed distribution W. Then (α_n) satisfies

$$\frac{\alpha_n - 2a_n}{b_n} \to \infty. \tag{9}$$

AF3 $K_n/n \to c$ such that $0 < c < 1 - t^\star$.

We have the following result

Theorem 2.2 *Let (u_n) be any positive and decreasing sequence such that $\sqrt{n}\, u_n \to \infty$. Then, under the asymptotic framework defined by* **AF1, AF2,AF3**,

$$P\left(\left|\frac{\hat{k}}{n} - t^\star\right| > u_n\right) \to 0. \tag{10}$$

Moreover, for $a > 0$ there exist constants c_1, c_2 which depend on a such that if

$$u_n = \frac{c_1 \alpha_n \sqrt{\log n}}{2\sqrt{n}} + \frac{c_2 \alpha_n \log(n)}{2n},$$

then

$$\mathbb{P}_{H_1(k_n^\star)}\left(|\frac{\hat{k}}{n} - t^\star| > u_n\right) \leq 2e^{-a\log(n)} + 2\mathbb{P}\left(\max_{1\leq i\leq n}|\varepsilon_i| > \alpha_n\right). \tag{11}$$

The proof of Theorem 2.2 is given in Section 3.

3 Proof of Theorem 2.2

Our procedure is based on two facts: a)under mild assumptions over the error distribution, if the null hypothesis is rejected, that is, if there is a group of significant coefficients and one of non significant coefficients, both groups of observations will be stochastically in order with high probability and b) for

two separate groups, separated at index $k_n^\star = [t^\star n]$, $T_{k_n,j} - Q_{k_n,j}$ will only converge at rate $\sqrt{n}$ for index k_n such that $|k_n - k_n^\star| = o(\sqrt{n})$.

Set $u_i = y_i$ for $i \in I_{k_n^\star}$ and $v_i = y_i$ for $i \notin I_{k_n^\star}$. Thus (v_i) is an i.i.d. sequence with distribution F_ε.

We have the following lemma that assures that both collections are stochastically in order with high probability:

Lemma 3.1 *Let $(u_{(i)})$ and $(v_{(i)})$ be the sequences $(|u_i|)$ and $(|v_i|)$ in a decreasing order. Then*

$$\mathbb{P}\big(v_{(1)} > u_{(k_n^\star)}\big) \to 0$$

and

$$\mathbb{P}\Big(v_{(1)} > \frac{\alpha_n}{2}\Big) \to 0$$

Proof: By assumption $(v_{(1)} - a_n)/b_n \xrightarrow{D} W$. On the other hand, let $(\tilde{v}_i)$ be a sequence of i.i.d. r.v. with distribution F_ε. Then,

$$\begin{aligned}
\mathbb{P}\big(v_{(1)} > u_{(k_n^\star)}\big) &\leq \mathbb{P}\big(v_{(1)} + \tilde{v}_{(1)} > \alpha_n\big) \\
&\leq \mathbb{P}\big(v_{(1)} > \alpha_n/2\big) + \mathbb{P}\big(\tilde{v}_{(1)} > \alpha_n/2\big) \\
&\leq 2\mathbb{P}\big(v_{(1)} > \alpha_n/2\big) \to 2W(\frac{\alpha_n/2 - a_n}{b_n}) \to 0. \quad \square
\end{aligned}$$

Lemma 3.1 yields the $(u_{(i)})$ and $(v_{(i)})$ are stochastically in order with high probability . Let Ω_n be the subset of Ω where $v_{(1)} < \alpha_n/2$ and $u_{(k_n^\star)} > \alpha_n/2$. Clearly $P(\Omega_n) \to 1$. In what follows we will restrict our proof to Ω_n.

We will denote $\mathbb{E}_k\,()$ the expectation under $\boldsymbol{H_1(k)}$ (instead of $\mathbb{E}_{H_1(k)}\,()$). On the other hand, let $a_i = \mathbb{E}_0\left(X_{(i)}\right) = \sum_{\ell=1}^n 1/\ell$.

1) Consider first the case $k > k_n^\star$. On Ω_n,

$$\begin{aligned}
T_{k,j} - Q_{k,j} &= T_{k,j} - B_{k,j,n}\, T_{k,K_n} \\
&= \big(T_{k,j} - \mathbb{E}_{k_n^\star}\,(T_{k,j})\big) - B_{k,j,n}\left(T_{(k,K_n} - \mathbb{E}_{k_n^\star}\,(T_{k,K_n})\right) \\
&\quad + \mathbb{E}_{k_n^\star}\,(T_{k,j}) - B_{k,j}\mathbb{E}_{k_n^\star}\,(T_{k,K_n}) \\
&= R_{k,j} + S_{k,j}
\end{aligned}$$

We have decomposed the statistics $T_{k,j} - Q_{k,j}$ into a random part $R_{k,j}$ and a deterministic part $S_{k,j}$. First, let $k = [tn]$ and $j = [sn]$ for $t \leq s$. As in Theorem 2.1, $R_{k,j}\mathbb{1}_{\Omega_n}$ (normalized by $\sqrt{n}$) as a process indexed by $(t,s) \in (0,1)^2$ converges in distribution to a zero-mean Gaussian process $\Gamma_{t,s} = (1-t^\star)[\Delta(s-t^\star) - \Delta(t-t^\star)]$.

On the other hand,

$$\begin{aligned}
S_{k,j} &= \mathbb{E}_{k_n^\star}\left(T_{k,j}\right) - \frac{\mathbb{E}_k\left(T_{k,j}\right)}{\mathbb{E}_k\left(T_{k,K_n}\right)}\,\mathbb{E}_{k_n^\star}\left(T_{k,K_n}\right) \\
&= \mathbb{E}_{k_n^\star}\left(T_{k,j}\right) - \mathbb{E}_k\left(T_{k,j}\right) - \frac{\mathbb{E}_k\left(T_{k,j}\right)}{\mathbb{E}_k\left(T_{k,K_n}\right)}\left(\mathbb{E}_{k_n^\star}\left(T_{k,K_n}\right) - \mathbb{E}_k\left(T_{k,K_n}\right)\right) \\
&= \sum_{i=j+1}^{j+k-k_n^\star} a_i - \sum_{i=1}^{k-k_n^\star} a_i + B_{k,j,n}\left(\sum_{i=K_n+1}^{K_n+k-k_n^\star} a_i - \sum_{i=1}^{k-k_n^\star} a_i\right) \\
&= \sum_{i=1}^{k-k_n^\star}\left(a_{i+j} - a_i + B_{k,j,n}(a_{i+K_n} - a_i)\right)
\end{aligned}$$

Thus, there exists a constant, $\gamma > 0$, which depends on c in [AF3], such that $\sup_j |S_{k,j}| \geq \gamma(k - k_n^\star)$ and

$$\begin{aligned}
&\mathbb{P}_{k_n^\star}\left(k_n^\star - \widehat{k} > n\,u_n\right) \leq \mathbb{P}\left(\eta_{k_n^\star} > \sup \eta_k\ ,\ (k - k_n^\star) > n\,u_n\right) \qquad (12)\\
\leq\ &\mathbb{P}\left(2\sup_k \sup_j R_{k,j} > \inf_k \sup_j |S_{k,j}|\ ,\ (k - k_n^\star) > n\,u_n\right) + \mathbb{P}(\Omega_n^c) \\
\leq\ &\mathbb{P}\left(2\sup_k \sup_j R_{k,j} > \gamma\, n\, u_n\right) + \mathbb{P}(\Omega_n^c)\,.
\end{aligned}$$

Because of the weak convergence of $R_{k,j}\mathbb{1}_{\Omega_n}$ the above probability tends to zero when n goes to infinity.

2) Consider now the case $k < k_n^\star$. On Ω_n,

$$\begin{aligned}
T_{k,j} - Q_{k,j} &= T_{k,j} - B_{k,j,n}\,T_{k,K_n} \\
&= (1 - B_{k,j,n})T_{k,k_n^\star - k} + \left(T_{k_n^\star,j} - \mathbb{E}\left(T_{k_n^\star,j}\right)\right) - B_{k,j,n}\left(T_{k_n^\star,K_n} - \mathbb{E}\left(T_{k_n^\star,K_n}\right)\right) \\
&\quad + \mathbb{E}\left(T_{k_n^\star,j}\right) - B_{k,j,n}\mathbb{E}\left(T_{k_n^\star,K_n}\right) \\
&= A_{k,j} + R_{k_n^\star,j} + U_{k,j}
\end{aligned}$$

where $A_{k,j} = (1 - B_{k,j,n})T_{k,k_n^\star - k}$. Remark that over Ω_n, $|y_{(i)}| > \alpha_n/2$. Then, there exists $c(\alpha_n) > 0$ such that $T_{k,k_n^\star - k} > c(\alpha_n)(k_n^\star - k)$, thus $A_{k,j} = \mathcal{O}(k_n^\star - k)$. On the other hand, $R_{k_n^\star,j}\mathbb{1}_{\Omega_n}$ converges in distribution to a zero-mean Gaussian process $\Gamma_{t^\star,s}$. Consider now the bias term $U_{k,j}$:

$$\begin{aligned}
U_{k,j} &= \mathbb{E}_{k_n^\star}\left(T_{k_n^\star,j}\right) - \frac{\mathbb{E}_k\left(T_{k,j}\right)}{\mathbb{E}_k\left(T_{k,K_n}\right)}\,\mathbb{E}_{k_n^\star}\left(T_{k_n^\star,K_n}\right) \\
&= \mathbb{E}_{k_n^\star}\left(T_{k_n^\star,j}\right) - \mathbb{E}_k\left(T_{k_n^\star,j}\right) - \frac{\mathbb{E}_k\left(T_{k_n^\star,j}\right)}{\mathbb{E}_k\left(T_{k,K_n}\right)}\left(\mathbb{E}_{k_n^\star}\left(T_{k_n^\star,K_n}\right) - \mathbb{E}_k\left(T_{k,K_n}\right)\right) \\
&\quad + \frac{\mathbb{E}_{k_n^\star}\left(T_{k_n^\star,K_n}\right)}{\mathbb{E}_k\left(T_{k,K_n}\right)}\,\mathbb{E}_k\left(T_{k,k_n^\star}\right) \\
&= \sum_{i=j+k-k_n^\star+1}^{j-k} a_i - \sum_{i=1}^{k_n^\star - k} a_i + \frac{\mathbb{E}_k\left(T_{k_n^\star,j}\right)}{\mathbb{E}_k\left(T_{k,K_n}\right)} \sum_{i=K_n+k-k_n^\star+1}^{K_n} a_i \ - \ \frac{\mathbb{E}_{k_n^\star}\left(T_{k_n^\star,K_n}\right)}{\mathbb{E}_k\left(T_{k,K_n}\right)} \sum_{i=1}^{k_n^\star - k} a_i.
\end{aligned}$$

Thus, there exists a constant $\delta > 0$, which depends on c in [AF3], such that $\sup_j |U_{k,j}| \geq \delta(k - k_n^\star)$ and $\mathbb{P}_{k_n^\star}\left(\widehat{k} - k_n^\star > n\,u_n\right) \to 0$ when $n \to \infty$. The latter together with (12) shows (10).

In order to show (11), sharper bounds on $\mathbb{P}_{k_n^\star}\big(\sup_k \sup_j R_{k,j} > C\, n\, u_n\big)$ are required for any given constant C. As above, we will restrict our attention to the set Ω_n and drop this fact from the notation. Consider first as above the case $k > k_n^\star$. Write, over Ω_n,

$$\begin{aligned} R_{k,j} &= \big(T_{k,j} - \mathbb{E}_{k_n^\star}(T_{k,j})\big) - B_{k,j,n}\left(T_{k,K_n} - \mathbb{E}_{k_n^\star}(T_{k,K_n})\right) \\ &= R_{k,j}^{(1)} + R_{k,j}^{(2)}. \end{aligned}$$

Remark $\sup_j B_{k,j,n} = 1$. So that $\sup_j |R_{k,j}^{(2)}| = |T_{k,K_n} - \mathbb{E}_{k_n^\star}(T_{k,K_n})|$.

Let G denote the common distribution function of the collection (X_i). We can rewrite

$$T_{k,K_n} - \mathbb{E}_{k_n^\star}(T_{k,K_n}) = \sum_i [X_i - \mathbb{E}_{k_n^\star}(X_i)]\mathbb{1}_{\{G^{-1}(1-K_n/n)<X_i\}}.$$

Thus,

$$\frac{T_{k,K_n} - \mathbb{E}_{k_n^\star}(T_{k,K_n})}{\alpha_n/2}$$

is the sum of independent bounded r.v. with variance bounded by 1, so that by Bennet's inequality

$$\begin{aligned} \mathbb{P}\left(\frac{\sup_j |R_{k,j}^{(2)}|}{\alpha_n/2} > \frac{\gamma/2nu_n}{\alpha_n/2}\right) &\leq \mathbb{P}\left(\frac{|T_{k,K_n} - \mathbb{E}_{k_n^\star}(T_{k,K_n})|}{\alpha_n/2} > \frac{c_1\gamma}{2\sqrt{2}}\sqrt{2n\log n} + \frac{3c_2\gamma}{2}\frac{\log n}{3}\right) \\ &\leq e^{-(a+1)\log(n)}, \end{aligned}$$

choosing $c_2 \geq \frac{2(a+1)}{3\gamma}$ and $c_1 \geq 2\frac{\sqrt{2}\sqrt{a+1}}{\gamma}$.

Hence, summing in k

$$\mathbb{P}\left(\sup_k \sup_{j<k+K_n} R_{k,j}^{(2)} > \frac{\gamma}{2nu_n}\right) \leq e^{-a\log n}.$$

For $R_{k,j}^{(1)}$ we have

$$\frac{T_{k,j} - \mathbb{E}_{k_n^\star}(T_{k,j})}{\alpha_n/2} = \sum_i [X_i - \mathbb{E}_{k_n^\star}(X_i)]\mathbb{1}_{\{G^{-1}(1-j/n)<X_i\}}.$$

Hence in this case we must use a functional version of Bennet's inequality (Theorem 7.3 in [4]) which yields

$$\mathbb{P}\left(\frac{\sup_j |R_{k,j}^{(2)}|}{\alpha_n/2} > \mathbb{E}\left(\frac{\sup_j |R_{k,j}^{(2)}|}{\alpha_n/2}\right) + \sqrt{2xv} + \frac{x}{3}\right) \leq e^{-x},$$

for $v \geq n + 2\mathbb{E}\left(\frac{\sup_j |R_{k,j}^{(2)}|}{\alpha_n/2}\right)$. Thus it remains to bound $A = \mathbb{E}\left(\frac{\sup_j |R_{k,j}^{(2)}|}{\alpha_n/2}\right)$. This can be done using standard symmetrization and entropy techniques to obtain, $A \leq 4\sqrt{n\log n}$, as the random entropy of the class $\mathcal{A} = \{\mathbb{1}_{G^{-1}(1-t),t\in[0,1]}\}$ (as it is a collection of increasing functions) is bounded by $2\log n$.

As above,

$$\begin{aligned}\mathbb{P}\left(\frac{\sup_j |R^{(1)}_{k,j}|}{\alpha_n/2} > \frac{\gamma/2nu_n}{\alpha_n/2}\right) &\leq \mathbb{P}\Big(\frac{2|T_{k,j} - \mathbb{E}_{k^\star_n}(T_{k,j})\,|}{\alpha_n} > 4\sqrt{n\log n} \\ &\quad + \sqrt{2(a+1)\log n(n+4\sqrt{n\log n})} + (a+1)\frac{\log n}{3}\Big) \\ &\leq e^{-(a+1)\log(n)},\end{aligned}$$

choosing $c_i, i = 1, 2$ appropriately.

The case $k < k^\star_n$ follows analogously.

4 Some extensions

4.1 Unknown distribution

Assume now that the distribution F_ε of the ε_i's is a parametric distribution $F_\varepsilon(\cdot\ ;\ \theta^\star)$, but where the parameter $\theta^\star$ is unknown. For any $0 \leq k \leq n-1$, let $\widehat{\theta}_k = \widehat{\theta}(y_{(k+1)}, y_{(k+2)}, \ldots, y_{(n)})$ be an estimator of θ. Let $F_{|\varepsilon|}(\cdot\ ;\ \theta^\star)$ be the distribution of the $|\varepsilon_i|$'s. We will consider the following assumptions:

F1 The cumulative distribution function $F_{|\varepsilon|}$ is two times differentiable as a function of θ with a.e. strictly positive derivative at $\theta = \theta^\star$.

F2 $\theta^\star$ belongs to some compact set Θ and there exists, under $H_{k^\star_n}$, a consistent estimator $\widehat{\theta}_{k^\star_n} = \widehat{\theta}(y_{(k^\star_n+1)}, y_{(k^\star_n+2)}, \ldots, y_{(n)})$ of $\theta^\star$.

F3 There exists (a, b) such that $0 < a < t^\star < b < 1$ and a Lipschitz continuous function $\tilde{\theta}$ defined on $[a, b]$ such that, under $H_{k^\star_n}$, $(\widehat{\theta}_{[tn]})$ converges uniformly on $[a, b]$ in probability to $(\tilde{\theta}(t))$.

Remark 1: under hypothesis **F2** and **F3**, $\widehat{\theta}_{k^\star_n}$ is a consistent estimator of $\tilde{\theta}(t^\star) = \theta^\star$.

Remark 2: When $t < t^\star$, convergence of $\widehat{\theta}_{[tn]}$ can be difficult to check with any estimator, since $\widehat{\theta}_{[tn]}$ depends on some y_i's that are not distributed under distribution F_ε. Nevertheless, it is possible to use an estimator based on some empirical quantiles and that only depends on the smallest observations, that is, that depends only on the observations distributed under F_ε.

For any $\theta \in \Theta$, let $X_i(\theta) = -\log\big(1 - F_{|\varepsilon|}(|y_i|, \theta)\big)$, and $T_{k,j}(\theta) = \sum_{i=k+1}^{k+j} X_{(i)}(\theta)$.

Then, we define the following procedure:

i) Let $K_n \leq [(1-b)\ n]$ be some positive integer. For $[a\ n] \leq k \leq n - K_n$,

   1. let $\widehat{\theta}_k = \widehat{\theta}(y_{(k+1)}, y_{(k+2)}, \ldots, y_{(n)})$,

2. for $i = 1, \ldots, n$, let $X_{(i)}(\widehat{\theta}_k) = -\log\Big(1 - F_{|\varepsilon|}(|y_{(i)}|; \widehat{\theta}_k)\Big)$,
3. for $1 \leq j \leq K_n$, compute

$$\begin{aligned} T_{k,j}(\widehat{\theta}_k) &= \sum_{i=k+1}^{k+j} X_{(i)}(\widehat{\theta}_k), \\ Q_{k,j}(\widehat{\theta}_k) &= B_{k,j,n}\, T_{k,K_n}(\widehat{\theta}_k), \\ \eta_k(\widehat{\theta}_k) &= \max_{k+1 \leq j \leq n} \frac{|T_{k,j}(\widehat{\theta}_k) - Q_{k,j}(\widehat{\theta}_k)|}{\sqrt{n}}. \end{aligned}$$

iii) Let

$$\hat{k} = \text{Arg} \min_{an \leq k \leq bn} \eta_k(\widehat{\theta}_k)$$

Remark: Here, $Q_{k,j}(\widehat{\theta}_k) = B_{k,j,n}\, T_{k,K_n}(\widehat{\theta}_k)$ is the conditional expectation of $T_{k,j}$, conditionally to T_{k,K_n}, assuming that $k_n^\star = k$ and that $\theta^\star = \widehat{\theta}_k$.

Then, we have the following result,

Theorem 4.1 *Assume* **F1, F2, F3**.

i) Introduce for $t \in [0,1]$ and $s \in [a,b]$ the random process

$$\hat{d}_n(t,s) = T_{k_n^\star,[K_n t]}(\widehat{\theta}_{[ns]}) - \mathbb{E}_{H_{k_n^\star}}\left(T_{k_n^\star,[K_n t]}(\widehat{\theta}_{[ns]}) | T_{k_n^\star,K_n}(\widehat{\theta}_{[ns]})\right).$$

Then, $\hat{d}_n(t,s)/\sqrt{n}$, as a stochastic process indexed on $[0,1] \times [a,b]$, converges in distribution, under $\boldsymbol{H_{k_n^\star}}$, to a zero mean Gaussian process $(\Lambda(t,s))$.

ii) Let (u_n) be any positive and decreasing sequence such that $\sqrt{n}\, u_n \to \infty$. Then, under the asymptotic framework defined by **AF1, AF2, AF3**,

$$\mathbb{P}_{H_1(k_n^\star)}\left(\left|\frac{\hat{k}}{n} - t^\star\right| > u_n\right) \to 0. \tag{13}$$

Proof: We first show *i)*.

With the above notation, for any $\theta \in \Theta$, let

$$\begin{aligned} \Psi_j(\theta) &= T_{k_n^\star,j}(\theta) - Q_{k_n^\star,j}(\theta) \\ \Psi_j'(\theta) &= \frac{\partial \Psi_j}{\partial \theta}(\theta) \end{aligned}$$

For any $t \in [0,1]$ and $s \in [a,b]$, let

$$\begin{aligned} \hat{d}_n(t,s) &= \Psi_{[nt]}(\widehat{\theta}_{[ns]}) \\ &= \Psi_{[nt]}(\tilde{\theta}(s)) + (\widehat{\theta}_{[ns]} - \tilde{\theta}(s))\Psi'_{[nt]}(\tilde{\theta}(s)) + +\mathcal{O}((\tilde{\theta}(s) - \widehat{\theta}_{[ns]})^2). \end{aligned}$$

Using the same proof used for the convergence of $(\Psi_{[nt]}(\theta^\star))/\sqrt{n}$ (see the Appendix), we show that, for any $s \in [a,b]$, $(\Psi_{[nt]}(\tilde{\theta}(s)))/\sqrt{n}$ and $(\Psi'_{[nt]}(\tilde{\theta}(s)))/\sqrt{n}$ also converge to two zero-mean Gaussian processes. Then, using hypothesis **F3**, $\widehat{\theta}_{[ns]} \to \tilde{\theta}(s)$ uniformly over $[a,b]$, and then, $(\hat{d}_n(t,s))$ converges to a zero mean Gaussian process $(\Lambda(t,s))$.

We show now *ii)*. For any $\theta \in \Theta$, let $a_i(\theta) = \mathbb{E}_{H_0}\left(X_{(i)}(\theta)\right)$. Following the proof of Theorem 2.2, consider first the case $k > k_n^\star$. On Ω_n, for any $\theta \in \Theta$,

$$\begin{aligned} T_{k,j}(\theta) - Q_{k,j}(\theta) &= \left(T_{k,j}(\theta) - \mathbb{E}_{k_n^\star}\left(T_{k,j}(\theta)\right)\right) - B_{k,j,n}\left(T_{k,K_n}(\theta) - \mathbb{E}_{k_n^\star}\left(T_{k,K_n}(\theta)\right)\right) \\ &\quad + \mathbb{E}_{k_n^\star}\left(T_{k,j}(\theta)\right) - B_{k,j}\mathbb{E}_{k_n^\star}\left(T_{k,K_n}(\theta)\right) \\ &= R_{k,j}(\theta) + S_{k,j}(\theta) \end{aligned}$$

As in Theorem 2.2, $R_{k,j}$ (normalized by $\sqrt{n}$) as a process indexed by $(t,s) \in (0,1)^2$ converges in distribution on Ω_n to a zero-mean Gaussian process $\Gamma_{t,s}(\theta)$.

On the other hand,

$$S_{k,j}(\theta) = \sum_{i=1}^{k-k_n^\star} \left(a_{i+j}(\theta) - a_i(\theta) + B_{k,j,n}(a_{i+K_n}(\theta) - a_i(\theta))\right)$$

Thus, there exists a constant, $\gamma > 0$, which depends on a, b in [F3], such that $\sup_j |S_{k,j}(\widehat{\theta}_k)| \geq (k - k_n^\star)\gamma$. We conclude that $\mathbb{P}_{k_n^\star}\left(k_n^\star - \widehat{k} > n\,u_n\right) \to 0$ using the arguments used for Theorem 2.2. The case $k < k_n^\star$ is identical. $\square$.

4.2 The unknown variance case

When $\theta^\star$ is a scale parameter, i.e. $F_\varepsilon(y;\theta^\star) = F_\varepsilon(y/\theta^\star;1)$, we introduce the following procedure which is scale invariant:

i) For $i = 1, \ldots, n$, let $X_{(i)} = |y_{(i)}|$,

ii) Let K_n be some positive integer. For $1 \leq k \leq n - K_n$ and $1 \leq j \leq K_n$, compute

$$T_{k,j} = \sum_{i=1}^{k+j} X_{(i)}, \tag{14}$$

$$Q_{k,j} = \frac{\mathbb{E}_{H_1(k)}\left(\sum_{i=k}^{k+j} X_{(i)}\right)}{\mathbb{E}_{H_1(k)}\left(\sum_{i=k}^{k+K_n} X_{(i)}\right)} T_{k,K_n}, \tag{15}$$

$$\eta_k = \max_{1 \leq j \leq K_n} \frac{|T_{k,j} - Q_{k,j}|}{\sqrt{n}}. \tag{16}$$

iii) Let

$$\hat{k}_u = \mathrm{Arg} \min_{1 \leq k \leq n - K_n} \eta_k$$

Remark: Notice, the minimization problem at hand is not changed if we consider $\frac{|T_{k,j}-Q_{k,j}|}{\sigma\sqrt{n}}$, so the procedure is indeed scale invariant. We have the following result, whose proof is omitted as it resembles quite closely that of Theorem 2.2.

Theorem 4.2 *Let (u_n) be any positive and decreasing sequence such that $\sqrt{n}\,u_n \to \infty$. Then, under the asymptotic framework defined by* **AF1, AF2,AF3**,

$$P(|\frac{\hat{k}_u}{n} - t^\star| > u_n) \to 0.$$

Moreover, for $a > 0$ there exist constants c_1, c_2 which depend on a such that if $u_n = \frac{c_1\alpha_n\sqrt{\log n}}{2\sqrt{n}} + \frac{c_2\alpha_n \log(n)}{2n}$ then

$$P(|\frac{\hat{k}_u}{n} - t^\star| > u_n) \leq 2e^{-a\log n} + 2P(\max_{1\leq i\leq n} \frac{|\varepsilon_i|}{\sigma} > \alpha_n).$$

4.3 Application to a regression problem

Consider as discussed in section 1 the following setting.

1. Assume we observe $y_i = f(x_i) + \varepsilon_i$ for a fixed collection x_i. Variables (ε_i) are assumed to be independent and identically distributed with variance $Var(\varepsilon) = \sigma^2$.
2. Associated to the collection (x_i), we introduce the empirical inner product $\langle t, s\rangle_n = \frac{1}{n}\sum_i t(x_i)s(x_i)$ and its associated empirical norm $\|\cdot\|_n$.
3. We are interested in approximating f in terms of a certain orthormal basis $\{\phi_\lambda\}_\lambda$. We assume that the basis is such that it is also orthonormal in the empirical norm $\langle , \rangle_n$.
4. Given the basis define the absolute empirical coefficients $\widehat{\beta}_j = \left|\langle y, \phi_j\rangle_n\right| \sqrt{n}$. More generally, we could consider the collection of the transformed coefficients $\gamma_j(h) = h(\widehat{\beta}_j/\sigma)$ for any given strictly increasing function h such that there exists β satisfying $h(ax) = a^\beta h(x)$ for any positive constant a.

In this section we are interested in the partial sums of the ordered variables $\widehat{\beta}$. If ε_i follows a Gaussian distribution then $y_j = \langle y, \phi_j\rangle_n \sqrt{n}/\sigma$ are independent normal variables and it is straightforward to show that the procedure considered in section 4.2 can be applied. More precisely assume the following assumptions are satisfied

R1 ε_i are an i.i.d. collection of centered normal r.v. with variance σ^2.

R2 $\{\phi_1, \ldots, \phi_n\}$ is orthonormal w.r.t. the empirical norm $\langle , \rangle_n$.

R3 For any $i \in I_{k_n^\star}$, $|\langle f, \phi_j\rangle_n| > a\sigma\sqrt{\log 2n}/\sqrt{n}$, with $a \geq 2\sqrt{2}$.

We have the following result,

Lemma 4.1 *Assume* **AF1, AF3, R1, R2** *and* **R3** *hold true. Let* $\hat{k}_u$*, be the estimator defined in section 4.2. Then, for* $b > 0$ *there exist constants* c_1, c_2 *which depend on* a *and* b *such that if* $u_n = \frac{c_1 \log(n)}{2\sqrt{n}} + \frac{c_2 \log^2(n)}{2n}$ *then*

$$P(|\frac{\hat{k}_u}{n} - t^\star| > u_n) \leq 2e^{-b \log n} + 2e^{-(a/2-\sqrt{2}) \log(n)}.$$

Proof: It follows directly from Theorem 4.2 by checking that if $\epsilon_i, i = 1, \ldots, n$ are independent standard normal random variables, then

$$P(\max_{1 \leq i \leq n} |\epsilon_i| > a\sqrt{2n}) \leq e^{-(a/2-\sqrt{2}) \log(n)}.$$

4.4 Random thresholding

It is interesting we can link this procedure to a random thresholding one, or as in [2, 3] in terms of penalized estimation. This link clearly appears when we use the ℓ_2-norm to define η_k:

$$\eta_k = n^{-2} \sum_{j=k+1}^{k+K_n} (T_{k,j} - Q_{k,j})^2$$

Lemma 2.2 ensures that good choice for the cutpoint between significant and non significant coefficients is $\widehat{k} = \arg\min \eta_k$. Thus, it is reasonable to assume we are looking from left to right to the first k such that $\eta_k > \eta_{k-1}$. We will assume coefficients are significant while $\eta_k < \eta_{k-1}$. In order to develop this idea we must understand how $\eta_k - \eta_{k-1}$ looks like. We have

$$\begin{aligned}
n^2(\eta_k - \eta_{k-1}) &= \sum_{j=1}^{K_n} (T_{k,j} - B_{k,j} T_{k,K_n})^2 - \sum_{j=1}^{K_n - 1} T_{k-1,j} - B_{k-1,j} T_{k-1,n})^2 \\
&= (X_k - B_{k-1,k} T_{k-1,k-1+K_n})^2 + \sum_{j=1}^{K_n-1} (T_{k,j} - B_{k,j} T_{k,K_n}))^2 \\
&\quad - \sum_{j=1}^{K_n-1} \Big(X_k + T_{k,j} - B_{k,j}(X_k + T_{k,K_n})(1 + o(1)) \Big)^2 \\
&\approx (X_k - B_{k-1,k} T_{k-1,k-1+K_n})^2 \\
&\quad + \sum_{j=1}^{K_n-1} X_k^2 (1 - B_{k,j,n})^2 + 2X_k(1 - B_{k,j,n})(T_{k,j} - B_{k,j,n} T_{k,k+K_n})
\end{aligned}$$

Hence coefficients will be significant approximatively until the first k such that

$$X_k \leq \tau_{k,n} := \frac{\sum_{j=1}^{K_n-1} (T_{k,j} - B_{k,j} T_{k,K_n})(1 - B_{k,j,n})}{\sum_{j=1}^{K_n-1} (1 - B_{k,j,n})^2}.$$

Remark: If using the estimator $\widehat{k}_u$, this would yield a scale free random estimator $\tau_{k,n}$ of the threshold τ. In the regression case, we obtain the traditional hard threshold scheme

$$\left|\langle y, \phi_j \rangle_n\right| > \frac{\tau_{k,n}}{\sqrt{n}}$$

5 Numerical experiments

We consider here the model

$$y_i = \mu_i + \varepsilon_i \tag{17}$$

where (ε_i) is a collection of i.i.d. r.v.

5.1 Testing the null hypothesis H_0

The distribution of $D_n = \max_j |T_j - \widehat{T}_j|/\sqrt{n}$ under H_0 is estimated by Monte-Carlo (using 5000 simulated samples). Here, the $(y_i; 1 \leq i \leq n)$ are i.i.d. $\mathcal{N}(0,1)$ r.v. We set $X_{(i)} = -\log(1 - F(y^2_{(i)}))$ where F is the cumulative distribution of a $\chi^2(1)$ distribution. Then, (T_j), $(\widehat{T}_j)$ and D_n are computed as described in Section 2.1.

Table 5.1 displays the estimated percentiles of order 0.50, 0.90, 0.95 and 0.99 obtained with different values of n. We see in this table that the distribution of D_n (except the tail) does not depend on n for $n \geq 20$. In particular, $\mathbb{P}_{H_0}(D_n > 0.65) \approx 0.05$ for any $n \geq 20$.

$n \setminus \alpha$	0.50	0.90	0.95	0.99
20	0.27	0.55	0.67	0.93
50	0.29	0.55	0.65	0.82
500	0.29	0.56	0.65	0.83
5000	0.30	0.55	0.64	0.79

Table 1: Estimated percentiles of D_n under H_0 obtained with different values of n

Using a level $\alpha = 5\%$, the test consist in rejecting the null hypothesis H_0 if $D_n > 0.65$. We estimated the power of this test, by simulating data under H_1. Here, the $(y_i; 1 \leq i \leq n/5)$ are i.i.d. $\mathcal{N}(\mu, 1)$ r.v. Figure 5.1 displays the estimated probability to reject the null hypothesis H_0 for different values of μ and n.

5.2 Estimating the number of significant coefficients

5.2.1 A Gaussian example

In the following experiment, we have simulated 500 Gaussian random variables, with $\mu_i = 4$ for $1 \leq i \leq 100$, and $\mu_i = 0$ for $101 \leq i \leq 500$. (ε_i) is a collection of $\mathcal{N}(0,1)$ i.i.d. r.v.

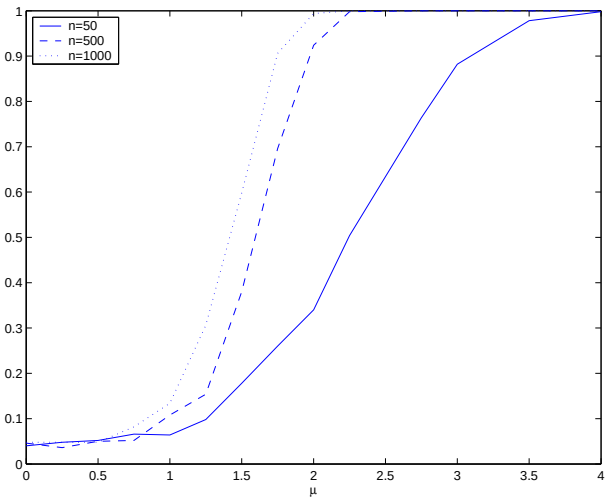

Figure 1: The estimated power of the 5% level test, for different values of μ and n.

Assuming that the variance of the ε_i's is known, we set

$$X_i = -\log(1 - F(y_i^2))$$

where F is the cumulative distribution function of a χ^2 r.v.

Then, we used the procedure described in Section 2.1. Figure 5.2.1 displays the two sequences (T_k) and (Q_k). We find $D_n = 14.95$ and reject the null hypothesis.

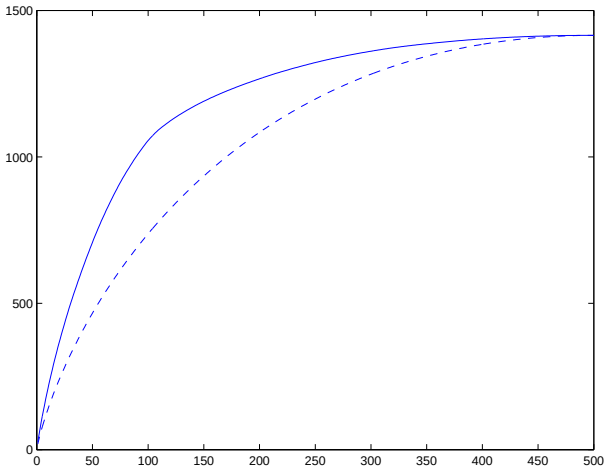

Figure 2: Example 1. The two sequences (T_k) and (Q_k)

After rejecting the null hypothesis, we will estimate the number of significant coefficients, following the procedure described in Section 2.2. We use here $K = 200$. Then, for $k = 1, 2, \ldots, 300$, we computed the sequences $(T_{k,j}, 1 \leq j \leq 200)$ and $(Q_{k,j}, 1 \leq j \leq 200)$. Figure 5.2.1 displays these two sequences for

$k = 70$, $k = 100$ and $k = 130$. We see that $(T_{k,j})$ concentrates around its conditional expected value $(\mathbb{E}_{H_1(k)}(T_{k,j}|T_{k,200}))$ only for $k = 100$. A bias is clearly present for $k = 70$ and $k = 130$. The sequence (η_k) defined by $\eta_k = \sum_{j=1}^{200}(T_{k,j} - Q_{k,j})^2/\sqrt{n-k}$ is displayed Figure 5.2.1. A minimum at $\hat{k} = 97$ is obvious.

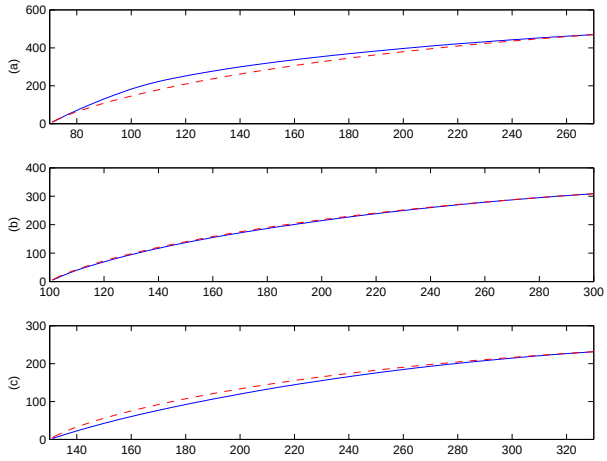

Figure 3: Example 1. The two sequences $(T_{k,j}, 1 \leq j \leq 200)$ and $(Q_{k,j}, 1 \leq j \leq 200)$ with (a) $k = 70$, (b) $k = 100$, (c) $k = 130$.

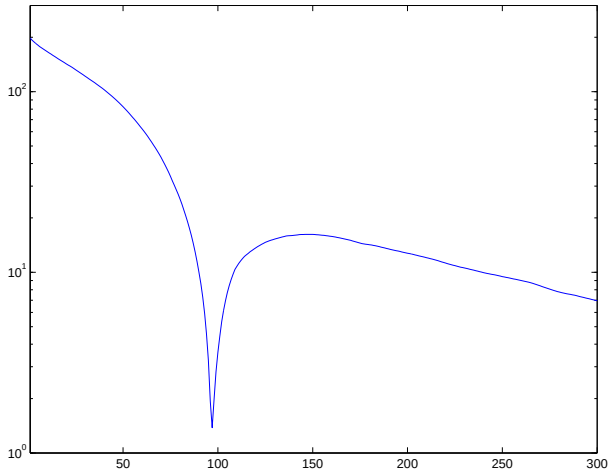

Figure 4: Example 1. The sequence (η_k) (in a semilog scale)

Repeating the same procedure with 100 simulated sequences, we obtained 100 values of $\hat{k}$. The mean value of $\hat{k}$ is 97.6 and the standard deviation is 4.8.

If we consider now that the variance is unknown, we use the procedure described Section 4.1, estimating the variance under $\mathbf{H}_1(k)$ by

$$\widehat{\theta_k} = \frac{1}{n-k} \sum_{i=k+1}^{n} y_{(i)}^2$$

The results obtained when the variance is unknown are very similar than those obtained when the variance is known. The mean value of $\hat{k}$ is 97.3 and the standard deviation is 5.1.

5.2.2 A Exponential example

In this second example, $n = 500$ again, but (ε_i) is a collection of *Expo(1)* i.i.d. r.v. Here, μ_i is uniformly distributed in $[3, 6]$ for $1 \leq i \leq 100$, and $\mu_i = 0$ for $101 \leq i \leq 500$.

When the parameter of the exponential distribution is known, we use the procedure described Section 2.1, setting $X_i = y_i$. Figure 5.2.2 displays the two sequences (T_k) and (Q_k). We find $D_n = 3.72$ and reject the null hypothesis.

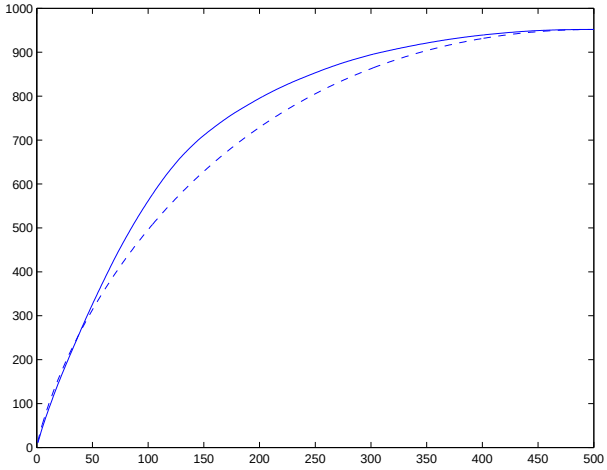

Figure 5: Example 2. The two sequences (T_k) and (Q_k)

The number of significant coefficients is estimated as before. Figure 5.2.2 displays these two sequences $(T_{k,j}, 1 \leq j \leq 200)$ and $(Q_{k,j}, 1 \leq j \leq 200)$ for $k = 70$, $k = 100$ and $k = 130$. In this example, the sequence (η_k) displayed Figure 5.2.2 is defined by $\eta_k = \sum_{j=1}^{200} |T_{k,j} - Q_{k,j}| / \sqrt{(n-k)^3}$. A minimum at $\hat{k} = 99$ is obvious.

Repeating the same procedure with 100 simulated sequences, we obtained 100 values of $\hat{k}$. The mean value of $\hat{k}$ is 103.8 and the standard deviation is 6.2.

The results obtained using the procedure described Section 4.1 when $\theta^\star$ is unknown are very similar: the mean value of $\hat{k}$ is 102.9 and the standard deviation is 5.6.

6 Appendix

Proof of proposition 2.1

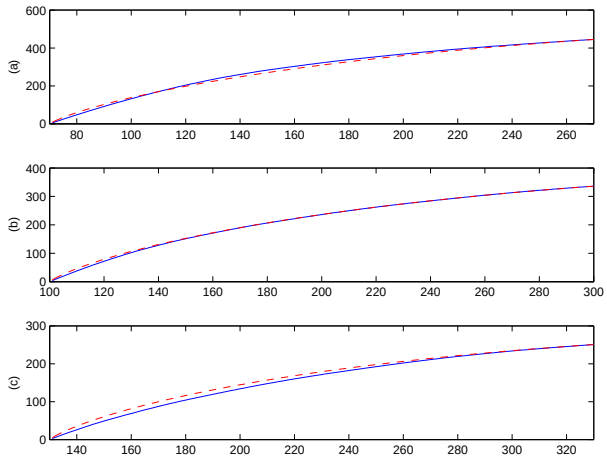

Figure 6: Example 2. The two sequences $(T_{k,j}, 1 \leq j \leq 200)$ and $(Q_{k,j}, 1 \leq j \leq 200)$ with (a) $k = 70$, (b) $k = 100$, (c) $k = 130$.

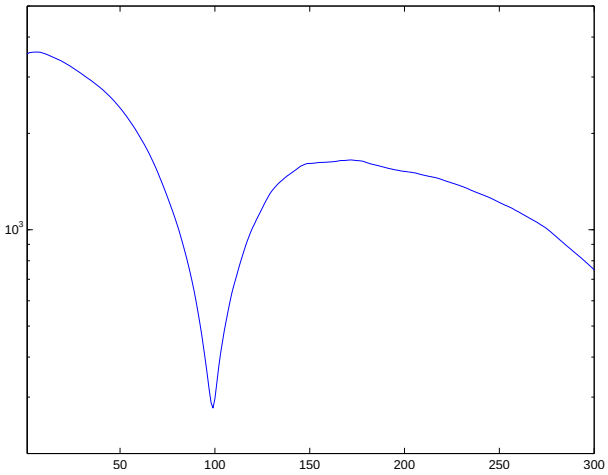

Figure 7: Example 2. The sequence (η_k) (in a semilog scale)

For any $1 \leq i \leq n$, let

$$D_i = X_i - X_{i+1}$$

with $X_{n+1} = 0$. Thus, $X_i = \sum_{j=i}^{n} D_j$. Next, let $Z_j = jD_j$. As is well known, $(Z_j; 1 \leq j \leq n)$ is a sequence of i.i.d random variables (*Exp(1)*), so that, for any $1 \leq k \leq K \leq n$,

$$\begin{aligned} \mathbb{E}\left(T_k | T_K\right) &= \sum_{i=1}^{k} \sum_{j=i}^{n} \mathbb{E}\left(D_j | T_K\right) \\ &= \left(\sum_{i=1}^{k} \sum_{j=i}^{n} \frac{1}{j} \right) \mathbb{E}\left(Z_1 | T_K\right) \end{aligned}$$

Since

$$\begin{aligned}\mathbb{E}\left(T_K|T_K\right) &= \left(\sum_{i=1}^{K}\sum_{j=i}^{n}\frac{1}{j}\right)\mathbb{E}\left(Z_1|T_K\right)\\ &= T_K\end{aligned}$$

and

$$\sum_{i=1}^{k}\sum_{j=i}^{n}\frac{1}{j} = k + k\sum_{j=k+1}^{n}\frac{1}{j}$$

we obtain

$$\mathbb{E}\left(T_k|T_K\right) = \frac{k + k\sum_{j=k+1}^{n}1/j}{K + K\sum_{j=K+1}^{n}1/j}\,T_K = \frac{B_{k,n}}{B_{K,n}}\,T_K. \quad \square \tag{18}$$

Proof of Theorem 2.1

Let $c_{nt,i} = \mathbb{1}_{[0,[nt]]}(i)$. By definition $\mathbb{E}\left(T_{[nt]}|T_n\right) = B_{[nt]}T_n$, so that $\mathbb{E}\left(T_{[nt]}\right) = \sum_i^n c_{nt,i}\mathbb{E}\left(X_{(i)}\right) = nB_{[nt]}$. Thus,

$$d_n(t) = \sum_i^n c_{nt,i}[X_{(i)} - \mathbb{E}\left(X_{(i)}\right)] - \frac{\sum_i^n c_{nt,i}\mathbb{E}\left(X_{(i)}\right)}{n}(T_n - n) = I_n(t) - II_n(t).$$

Let $G_n = \sum_{i=1}^n \zeta_{n-i}$ stand for the empirical sum of uniform r.v. ζ_i. Then, as in [9] it can be seen that

$$\frac{1}{\sqrt{n}}I_n(t) = -\frac{1}{\sqrt{n}}\int_0^1 [G_n - I](s)\mathbb{1}_{[0,t]}(s)dF^{-1}(s) + o_p(1)$$

and

$$\frac{1}{\sqrt{n}}II_n(t) = -(t - t\log(t))\frac{1}{\sqrt{n}}\int_0^1 [G_n - I](s)dF^{-1}(s) + o_p(1),$$

where $o_p(1)$ is uniform for all $t \in [0,1)$. So that,

$$\frac{1}{\sqrt{n}}d_n(t) = \int_0^1 [R^t(u) - (t - t\log(t))F^{-1}(u)]d[G_n(s) - (1-s)] + o_p(1),$$

with $R^t(u) = \int_0^u dF^{-1}(s)\mathbb{1}_{[0,t]}(s)ds$. The result now follows because

$$\mathcal{G} = \{R^t - (t - t\log(t))F^{-1}, t \in [0,1]\}$$

is a Donsker class.

Acknowledgments: We thank very much José Rafael León, Jean-Michel Loubes and Pascal Massart for stimulating discussions.

References

[1] B. Arnold, N. Balakrishnan, and H. Nagaraja. *A first course in order statistics*. Wiley series in probability, 1993.

[2] A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413, 1999.

[3] L. Birgé and P. Massart. Minimal penalties for gaussian model selection. *Probab. Theory Related Fields (to appear)*, 2005.

[4] O. Bousquet. Concentration inequalities for sub-additive functions using the entropy method. In *Stochastic inequalities and applications*, volume 56 of *Progr. Probab.*, pages 213–247. Birkhäuser, Basel, 2003.

[5] D. Donoho and I. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81:425–455, 1994.

[6] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96(456):1348–1360, 2001.

[7] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer, series in statistics, 2001.

[8] M. S. Pinsker. Optimal filtration of square-integrable signals in Gaussian noise. *Probl. Peredachi Inform.*, 2(16):52–68, 1980.

[9] G. Shorak and J. Wellner. *Empirical processes with Applications to Statistics*. Wiley, 1986.

[10] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal Statist. Soc. B.*, 58:267–288, 1996.

Contents

Unité de recherche INRIA Futurs
Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)

Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399